This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

LLNL-PRES-712737

State of the Earth System Grid Federation (ESGF)

ESGF Earth System Grid Federation

http://esgf.llnl.gov

6th Annual Earth System Grid Federation
2016 Face-to-Face Conference

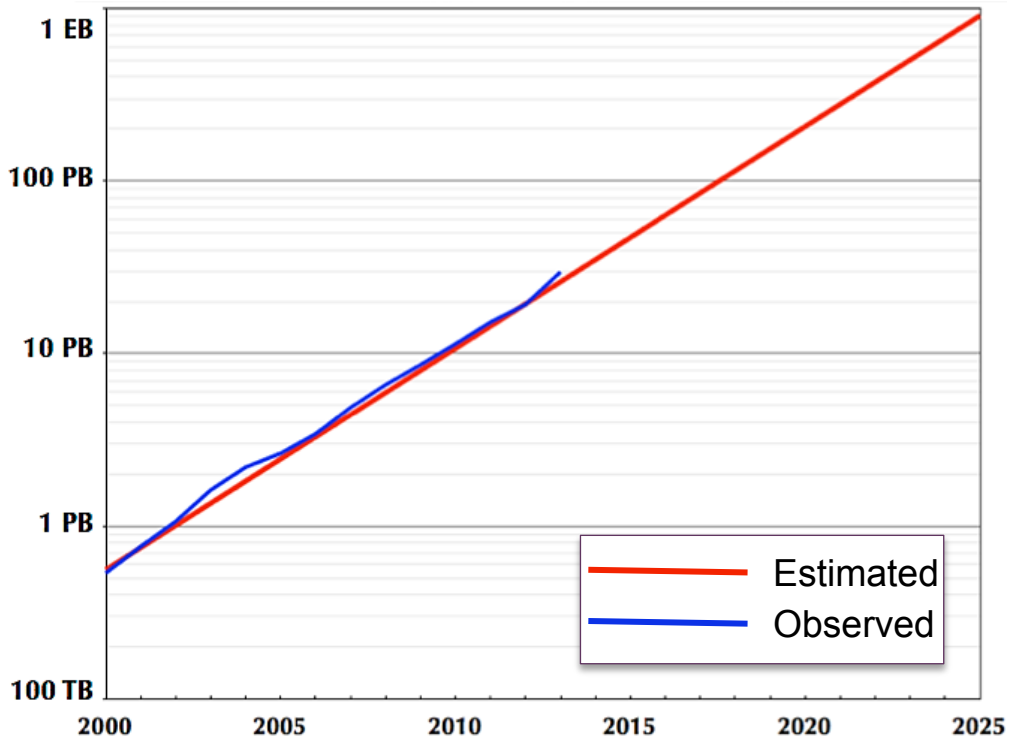
ESGF Executive Committee and Conference Organizers

Dean N. Williams (Chair, DOE)
Michael Lautenschlager (Co-Chair, DKRZ)
Luca Cinquini (NASA/NOAA)
Sébastien Denvil (IPSL)
Robert Ferraro (NASA)
Daniel Duffy (NASA)
V. Balaji (NOAA)
Claire Trenham (NCI)

Automating infrastructure for archiving and comparing simulation/observation results, diagnostics, and validations

- Leading the creation of a flexible, extensible infrastructure for future national and international climate efforts
- Exploiting broad data sharing for scientific breakthroughs
- Projects represent state-of-the-art in several disciplines
- Petabytes of data distributed around the world
- Requires combining model output with various kinds of observations and reanalysis
- Data and algorithms must be managed, tracked and validated
- Empowering advances by integrating our high-quality data streams for a virtual laboratory

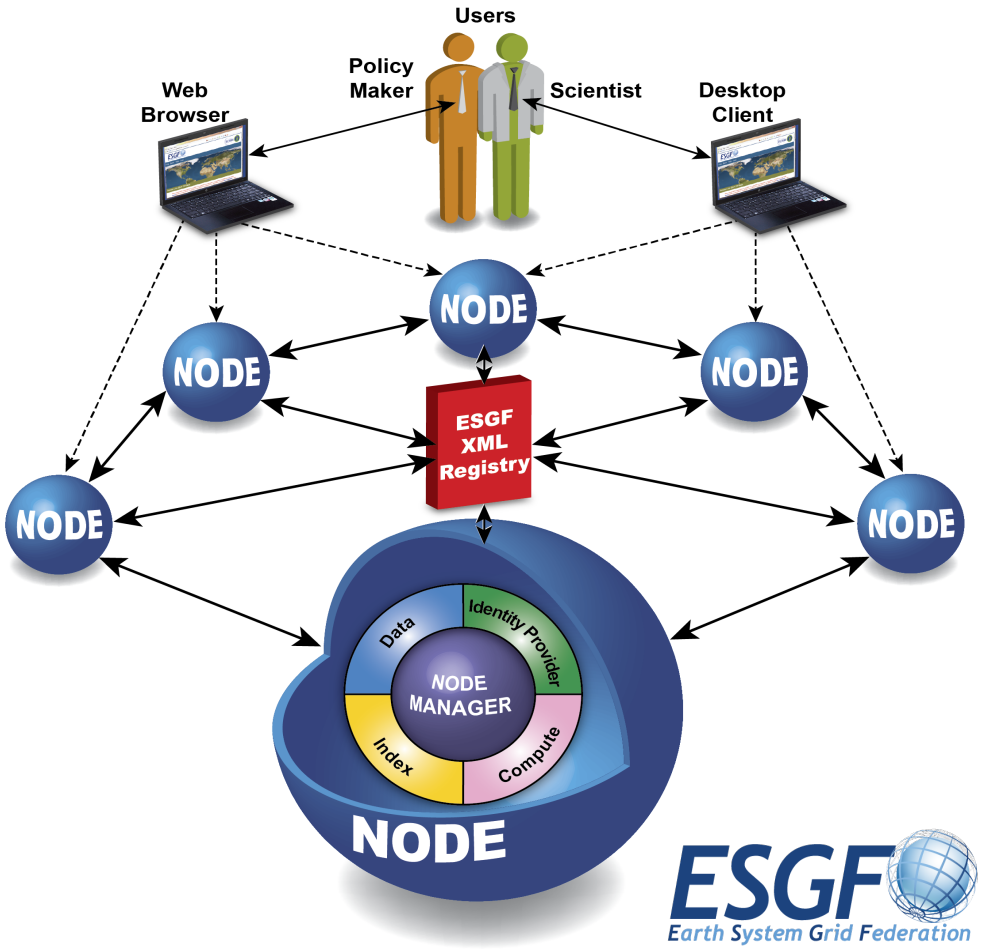
Community Earth System Model (CESM) archive volume image courtesy of Gary Strand



ESGF is leading the climate community in integrating all existing and future data holdings into a seamless and unified environment. We are working on new methods to deal with rapid data growth.

Federated distributed data archival and retrieval system

- Federated peer-to-peer architecture
- Support discipline-specific portals
- Support browser-based and direct client access
- Single sign-on
- Automated script and GUI-based publication tools
- Full support for data aggregations
- "Collaborative Documentation for Scientific Projects" (i.e., CoG)



Enabling climate research in a data-rich environment.

Collaboration between interagency partners from disparate domains enabled development of the ESGF[1] data ecosystem

Agency Sponsors

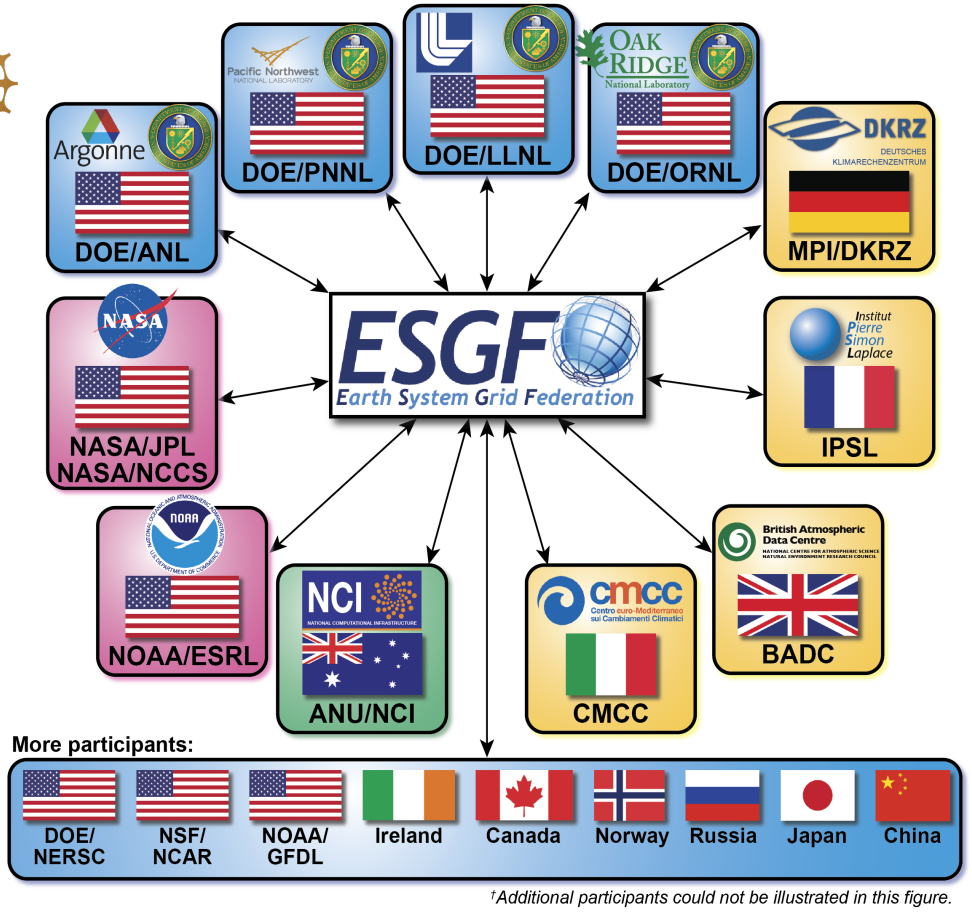
DOE IS-ENES NASA NCI NOAA NSF

ESGF Steering Committee:

- Justin Hnilo (Chair, DOE), Sylvie Joussaume (IS-ENES), Tsengdar Lee (NASA), Ben Evans (NCI), Annarita Mariotti (NOAA)

ESGF Executive Committee:

- Dean N. Williams (Chair, DOE), Michael Lautenschlager (co-Chair, DKRZ), Luca Cinquini (NASA/NOAA), Sébastien Denvil (IPSL), Robert Ferraro (NASA), Daniel Duffy (NASA), V. Balaji (NOAA), Claire Trenham (NCI)



†Additional participants could not be illustrated in this figure.

1. ESGF Executive Committee and Working Team Leads, 2015 5th Earth System Grid Federation Annual Report, DOI: 10.2172/1253685.

2016 ESGF user survey results

Which option describes you?	Responses	
Data Provider	7.95%	29
Data Consumer	63.84%	233
Both Provider and Consumer	28.22%	103
Total	365	

Which option describes you?	Responses	
Undergraduate Student	2.82%	9
Graduate Student	15.05%	48
Post-Doc	22.57%	72
Academic Scientist/Professional	31.66%	101
Government Scientist/Professional	22.88%	73
Private Scientist/Professional	2.19%	7
Responses Other (please specify)	2.82%	9
Total	319	

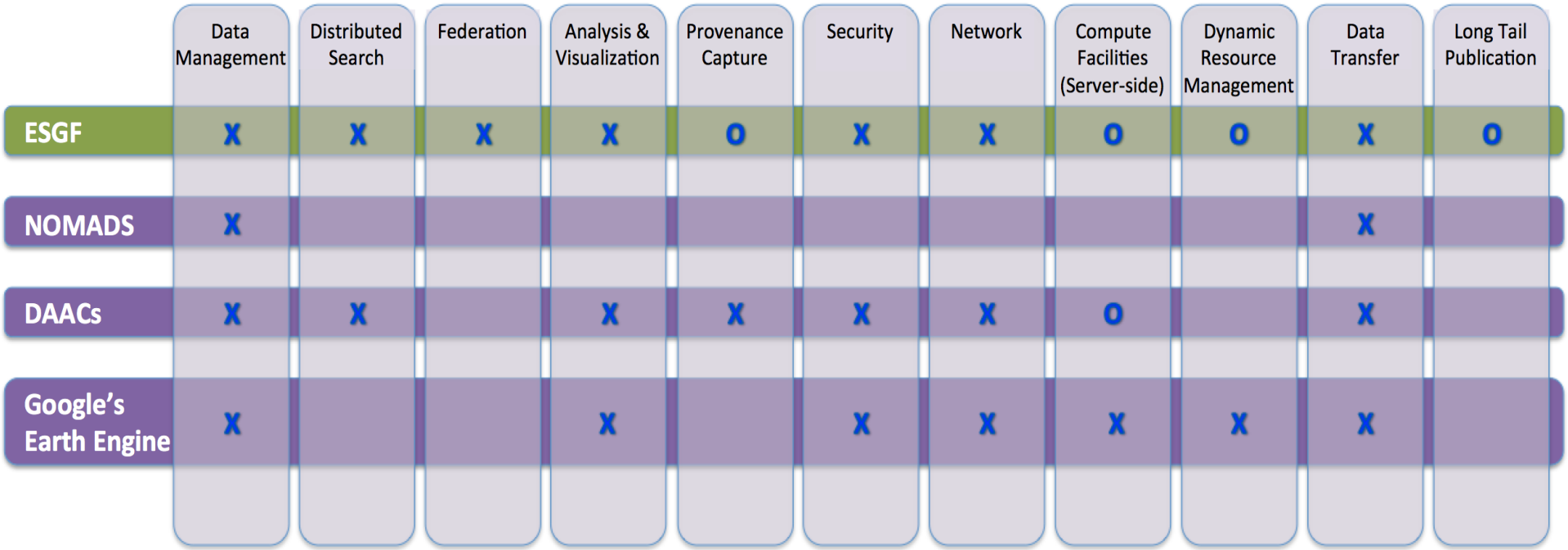
Affiliation?	Responses	
Government Agency	37.62%	120
University	57.37%	183
Private Sector	5.02%	16
Total	319	

*Top needs identified by the survey. Survey question types: **red** question text indicates needed ESGF improvements; **blue** question text indicates needed capabilities identified by the community; and **green** question text indicates features that the community finds most useful.*

ESGF
Earth System Grid Federation

Question	Category	Total Response	Weighted Score	Percentage of Response	Combined Weighted Score and Percentage of Response
Which feature of ESGF do you find most difficult to use, and/or you think needs the most improvement?	User Interface (the web sites or "CoG")	135	3.78	39.36%	1.49
How important is knowledge gathering, managing, and sharing?	Ingest and access to large volumes of scientific data (i.e., from data archive to super computer and server-side analysis)	119	4.22	34.69%	1.46
Which feature of ESGF do you find most difficult to use, and/or you think needs the most improvement?	Web documentation	129	3.89	37.61%	1.46
How good are human–computer interactions?	Improved designs and principles of user interfaces to enable easier access to computer and software capabilities (e.g. recommendation systems, more flexible and interactive interfaces)	106	4.13	30.90%	1.28
Which feature of ESGF do you find most difficult to use, and/or you think needs the most improvement?	Distributed global search	118	3.67	34.40%	1.26
How important is knowledge gathering, managing, and sharing?	Quality control algorithms for data	110	3.86	32.07%	1.24
How useful is user support?	Data access and usage	101	4.21	29.45%	1.24
Which features of ESGF do you find most useful?	User interface (the web sites, or "CoG")	100	4.22	29.15%	1.23
Which features of ESGF do you find most useful?	Distributed global search	91	4.55	26.53%	1.21
Is resource management needed?	Reliability and resilience of resources	97	4.25	28.28%	1.20
How important is knowledge gathering, managing, and sharing?	Interoperability: Interfaces that ensure a high degree of interoperability at format and semantic level between repositories and applications	107	3.77	31.20%	1.17
How important is rapid information retrieval, knowledge-based response, and decision-making mechanisms?	Availability of ancillary data products such as data plots, statistical summaries, data quality information, and other documents	104	3.84	30.32%	1.16
Is resource management needed?	Access to sufficient observational and experimental resources	99	4.01	28.86%	1.16
Is resource management needed?	Awareness and information of availability of these resources	98	4.04	28.57%	1.15
Is resource management needed?	Access to enough computational and storage resources	99	3.89	28.96%	1.12
How good are human–computer interactions?	Environments that support more effective collaboration and sharing within and between science teams (e.g., collaboration tools)	105	3.63	30.61%	1.11
How useful is user support?	Data sharing	96	3.96	27.99%	1.11
Which feature of ESGF do you find most difficult to use, and/or you think needs the most improvement?	Direct data delivery into ESGF computing systems from distributed data resources	95	3.99	27.70%	1.10
How important is rapid information retrieval, knowledge-based response, and decision-making mechanisms?	Reproducibility	106	3.57	30.90%	1.10

ESGF comparison to other archives



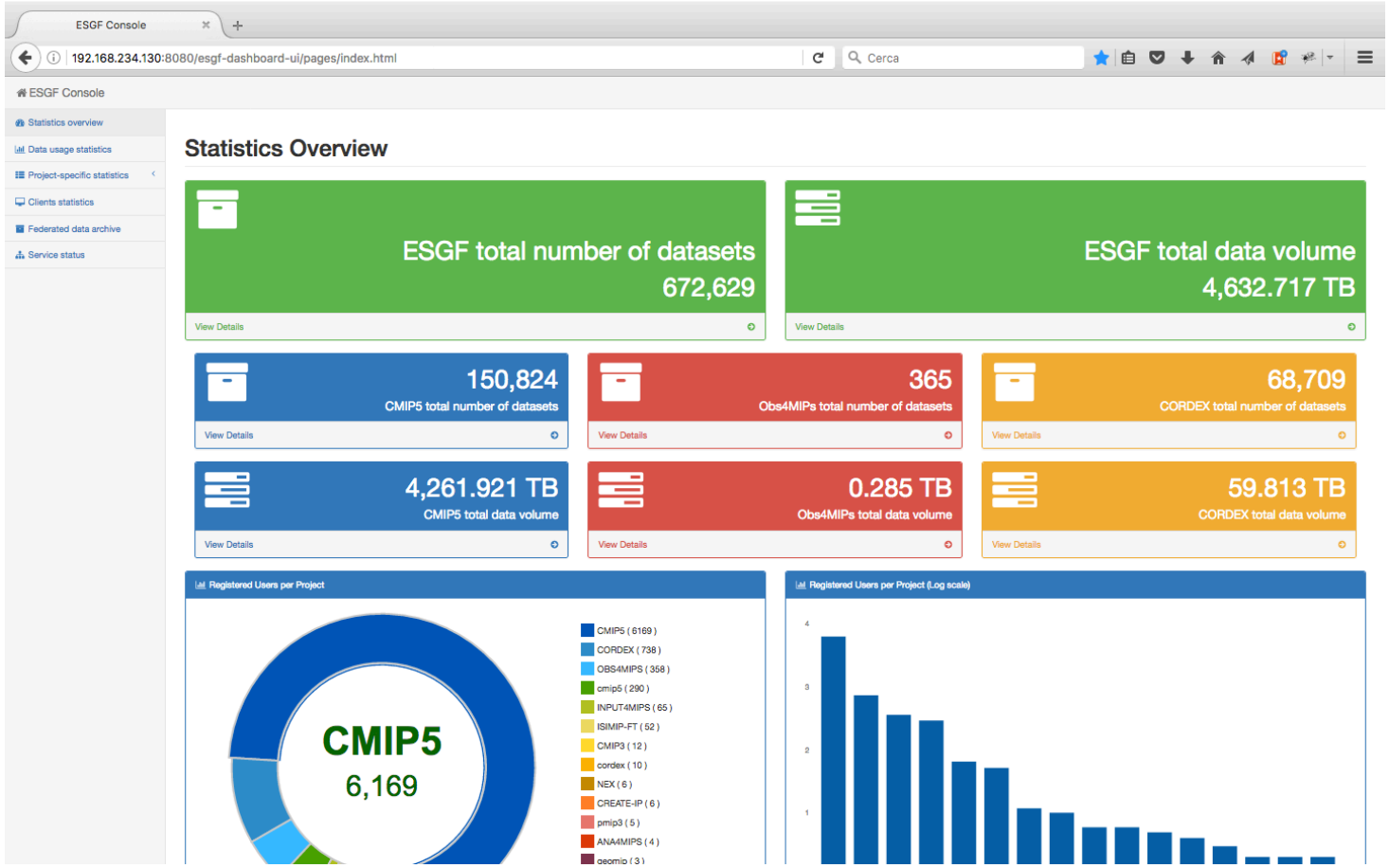
	Data Management	Distributed Search	Federation	Analysis & Visualization	Provenance Capture	Security	Network	Compute Facilities (Server-side)	Dynamic Resource Management	Data Transfer	Long Tail Publication
ESGF	X	X	X	X	O	X	X	O	O	X	O
NOMADS	X									X	
DAACs	X	X		X	X	X	X	O		X	
Google's Earth Engine	X			X		X	X	X	X	X	

X **Current capabilities** O **Future capabilities**

ESGF is the only archive that allows interoperability among disparate data sets (i.e., simulations, observations, and reanalysis data) for assessment study.

2016 ESGF user/usage demographics: statistics overview

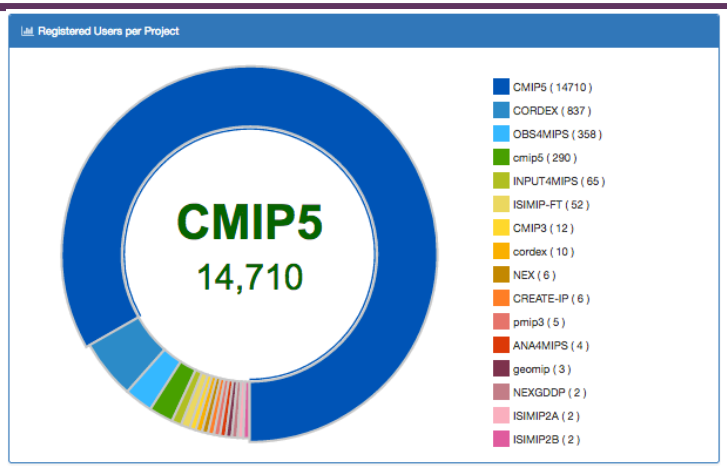
- ESGF supported over 672,000 datasets from universities, national and international laboratories and manages more than 4.6 petabytes of data.

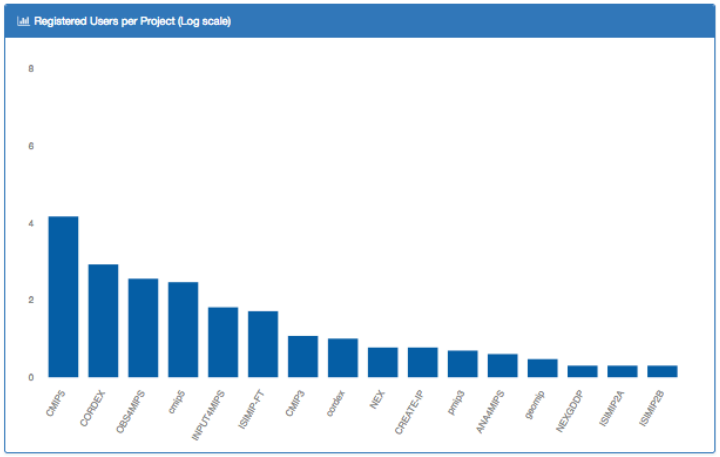


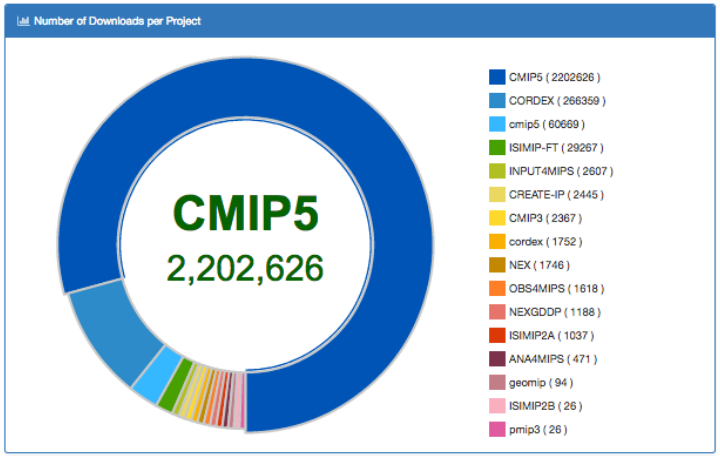
ESGF usage demographics can be access via a dashboard or desktop.

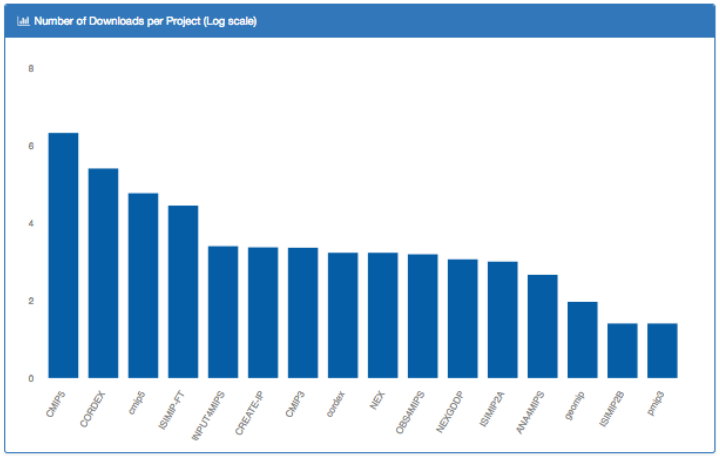
2016 ESGF user/usage demographics: users per project and number of downloads

- ESGF supported over 16,000 active users from universities, national and international laboratories and millions of data downloads.



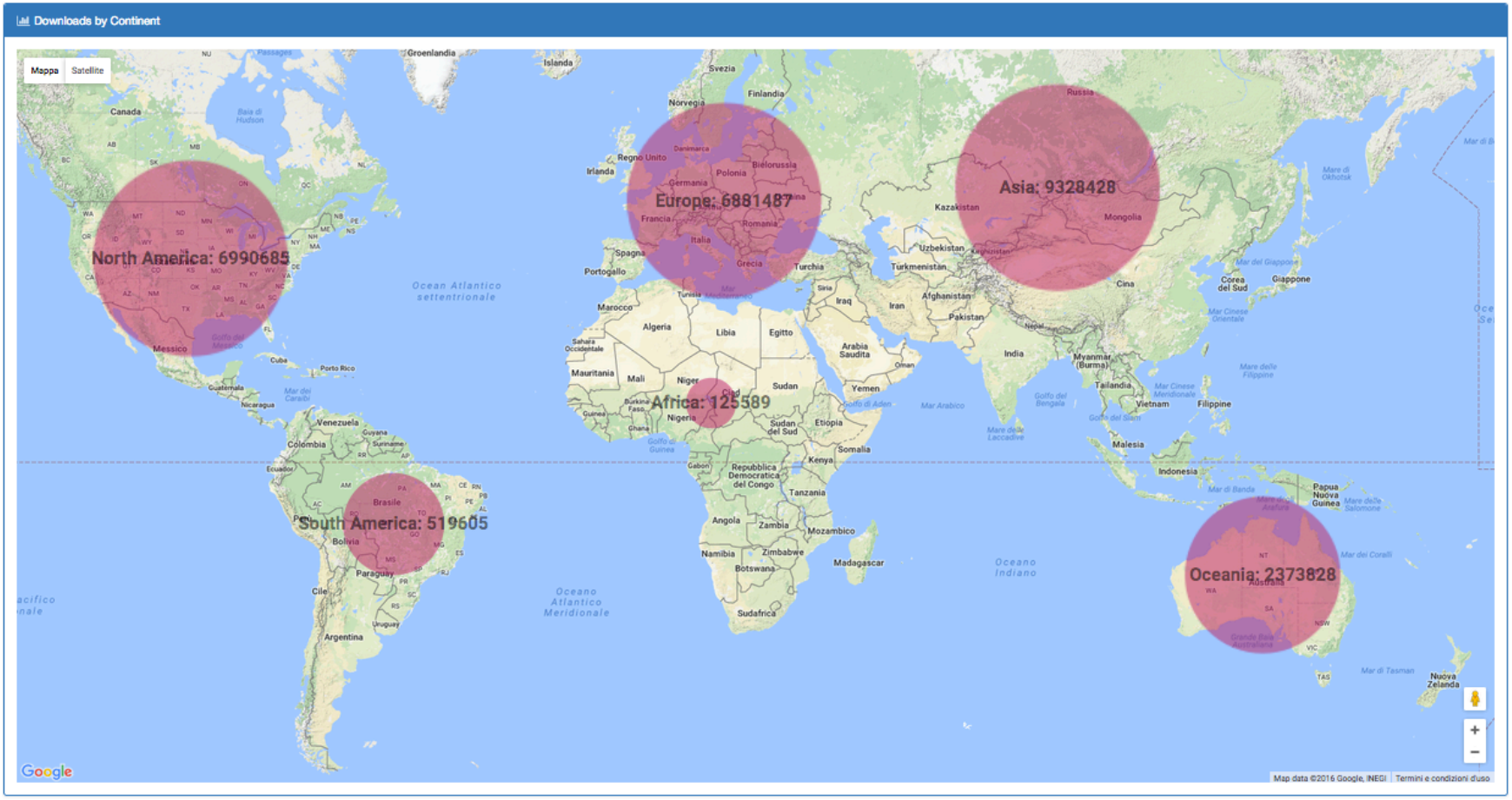






Not all ESGF projects and node sites are represented in the display.

2016 ESGF user/usage demographics: downloads by continents

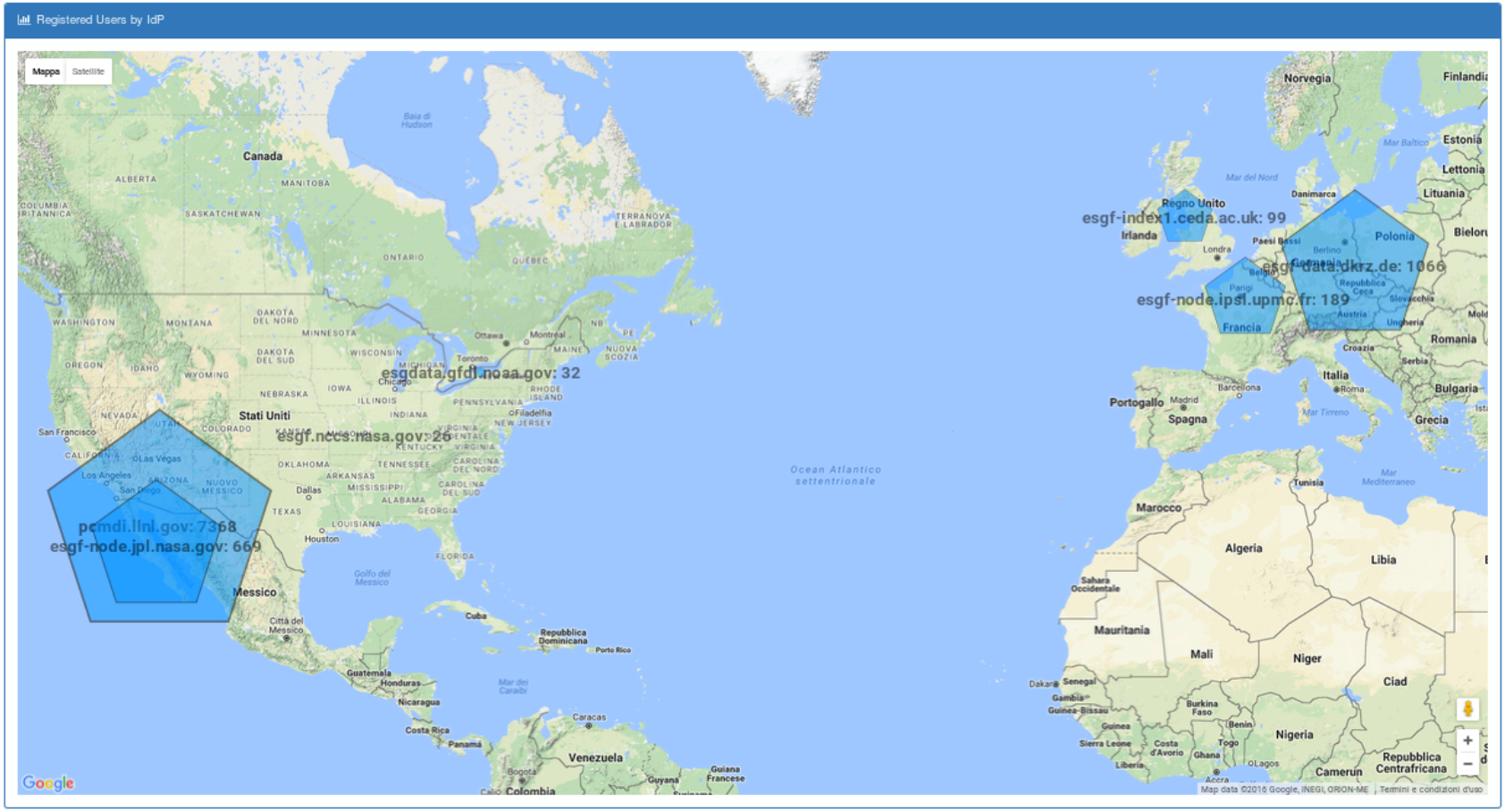


Not all ESGF projects and node sites are represented in the display.

2016 ESGF user/usage demographics: users by continents

Not all ESGF projects and node sites are represented in the display.

2016 ESGF user/usage demographics: registered users by specific sites



Not all ESGF projects and node sites are represented in the display.

ESGF archives are important to climate research science missions

Table 1. ESGF Index Nodes

Institute	Gateway URL	Version	Country	Project	Contact
CEDA	https://esgf-index1.ceda.ac.uk/	2.2.3	U.K.	**CMIP5, CORDEX, Obs4MIPs, SPECS, ESA CCI, EUCLEIA, CLIPC**	alan.iwi@stfc.ac.uk
DKRZ	https://esgf-data.dkrz.de/	2.3.8	Germany	**CMIP5, CORDEX, Obs4MIPs, ISI-MIP**	berger@dkrz.de
ESRL/NOAA	https://esgf.esrl.noaa.gov/	2.3.8	U.S.	**HIWPP, Coupled NEMS**	sylvia.murphy@noaa.gov
GFDL/NOAA	https://esgdata.gfdl.noaa.gov/	2.2.3	U.S.	**CMIP5, ncpp2013, Obs4MIPs**	hans.vahlenkamp@noaa.gov
GSFC/NASA	https://esgf.nccs.nasa.gov/	2.3.8	U.S.	**CMIP5, Obs4MIPs, Ana4MIPs, NEX-GDDP, NEX-DCP30, CREATE-IP**	daniel.q.duffy@nasa.gov
IPSL	https://esgf-node.ipsl.upmc.fr/	2.2.3	France	**CMIP5, CORDEX, Obs4MIPs**	sebastien.denvil@ipsl.jussieu.fr
JPL/NASA	https://esgf-node.jpl.nasa.gov/	2.3.8	U.S.	**Obs4MIPs, GASS-YoTC, CMAC**	Luca.Cinquini@jpl.nasa.gov
LLNL	https://pcmdi.llnl.gov/	2.3.8	U.S.	**CMIP5, CMIP3, input4MIPS, ACME**	sasha@llnl.gov
LiU	https://esg-dn1.nsc.liu.se/	2.2.2	Sweden	**CMIP5, CORDEX, Obs4MIPs**	pchengi@nsc.liu.se

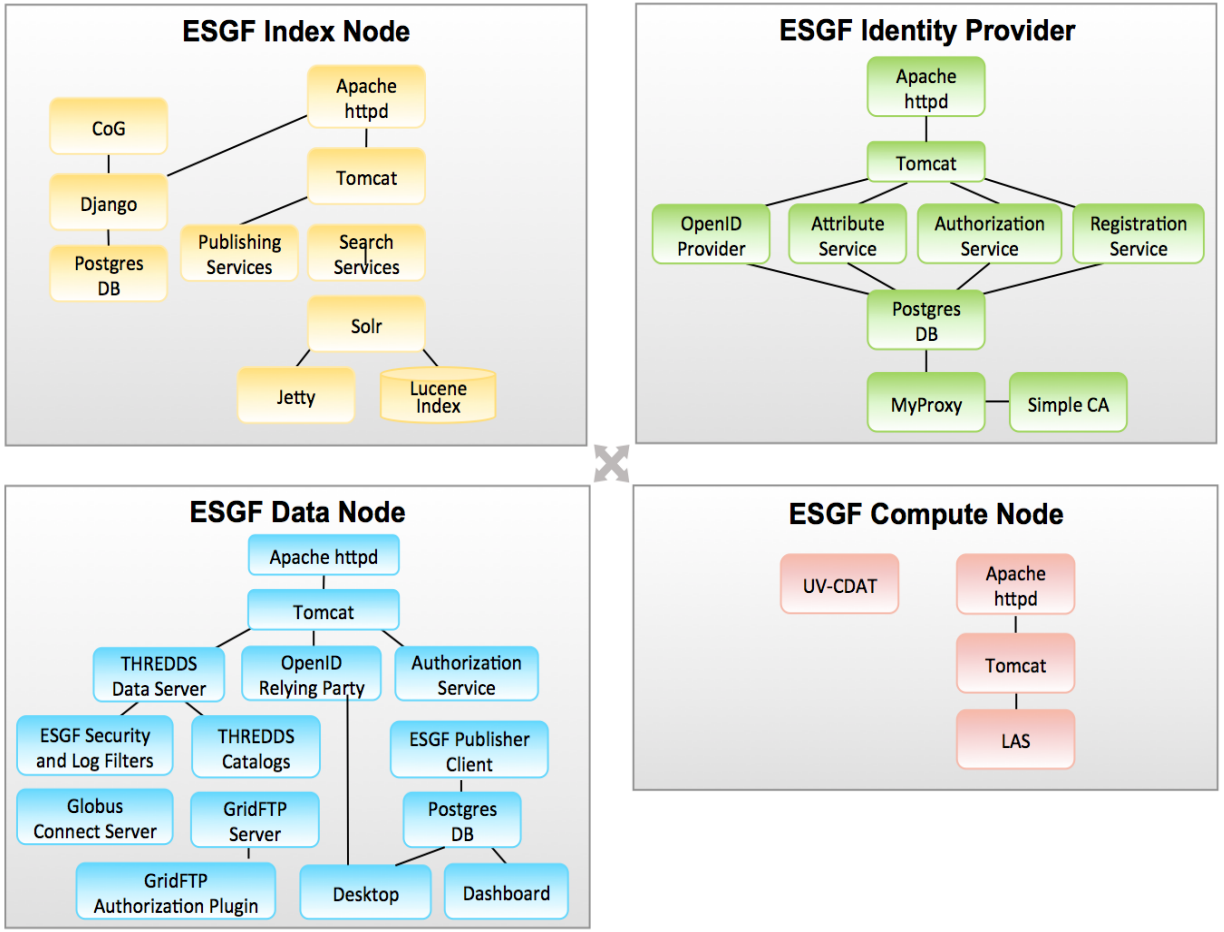
Table 2. ESGF Data Nodes

Data Node	Number of Datasets
aims3.llnl.gov	**30711**
cordexesg.dmi.dk	**11457**
eridanus.eoc.dlr.de	**1**
esg-	**558**
esg-dn1.nsc.liu.se	**61142**
esg.ccs.ornl.gov	**23**
esg.cnrm-game-	**1497**
esgdata.gfdl.noaa.gov	**5693**
esgf-data.jpl.nasa.gov	**16**
esgf-data1.ceda.ac.uk	**14808**
esgf-data1.diasjp.net	**8308**
esgf-data2.ceda.ac.uk	**386955**
esgf.extra.cea.fr	**3974**
esgf.ichec.ie	**961**
esgf.knmi.nl	**1**
esgf.nccs.nasa.gov	**4803**
esgf1.dkrz.de	**16740**
noresg.norstore.no	**911**
plot1.ornl.gov	**261**
prodn.idris.fr	**120**

Tables 1 and 2 above show ESGF global production gateways and data nodes in use, providing data to the greater community.

ESGF immediate challenges

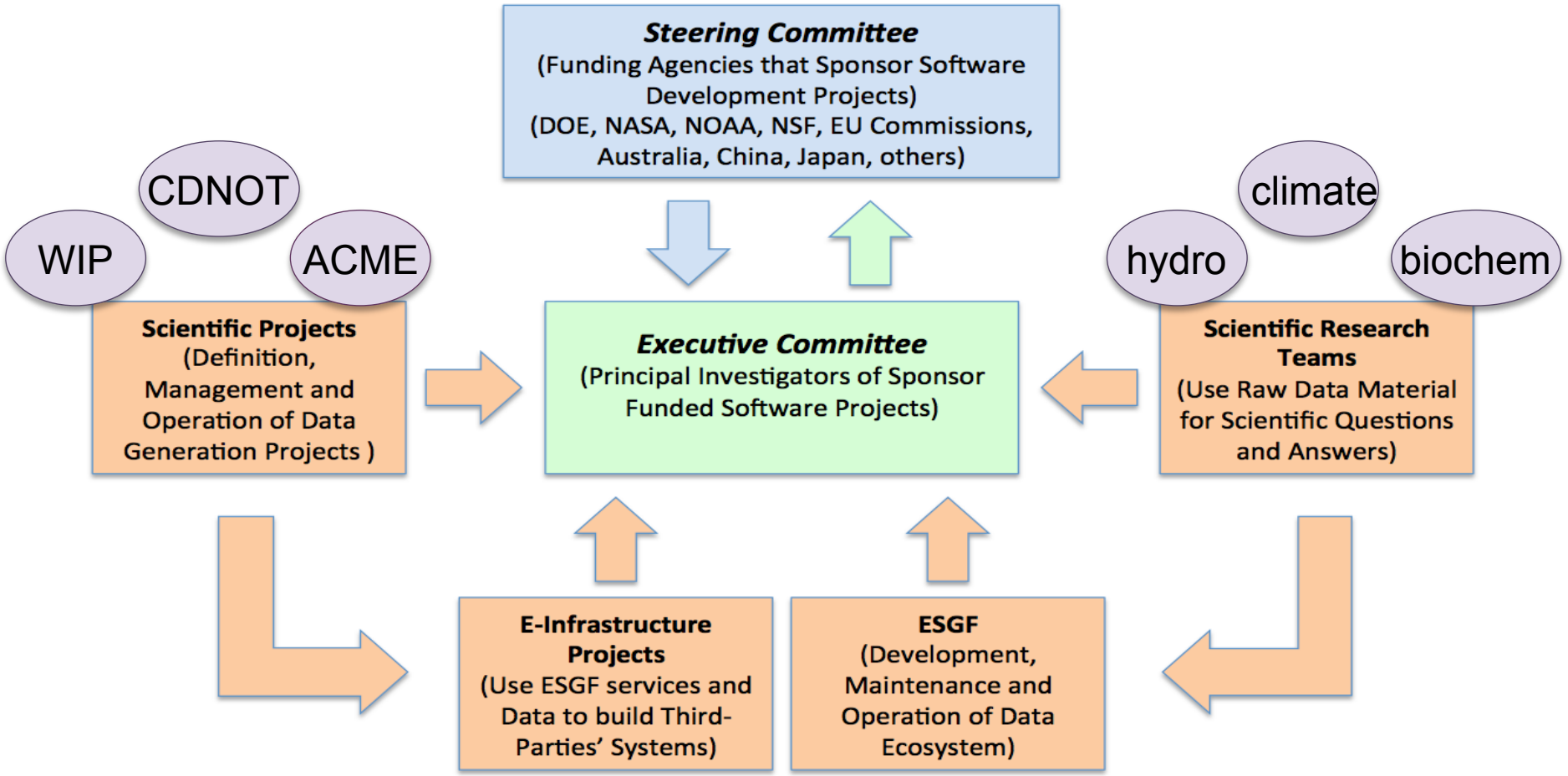
- CMIP6 and all the MIPs
- Increasingly serving observations alongside models, as equal citizens
- Enabling server-side computation
- Integration with other systems, such as DAACs, ACME, Copernicus
- Scaling - which requires new architectural approaches and more modularization and resource management
- Making installation much easier



The software components that constitute the ESGF Node software stack are logically grouped by "flavor" or area of functionality. Each flavor can be installed separately, and flavors can be deployed in multiple combinations.

ESGF management and team integration

Governance communication architecture



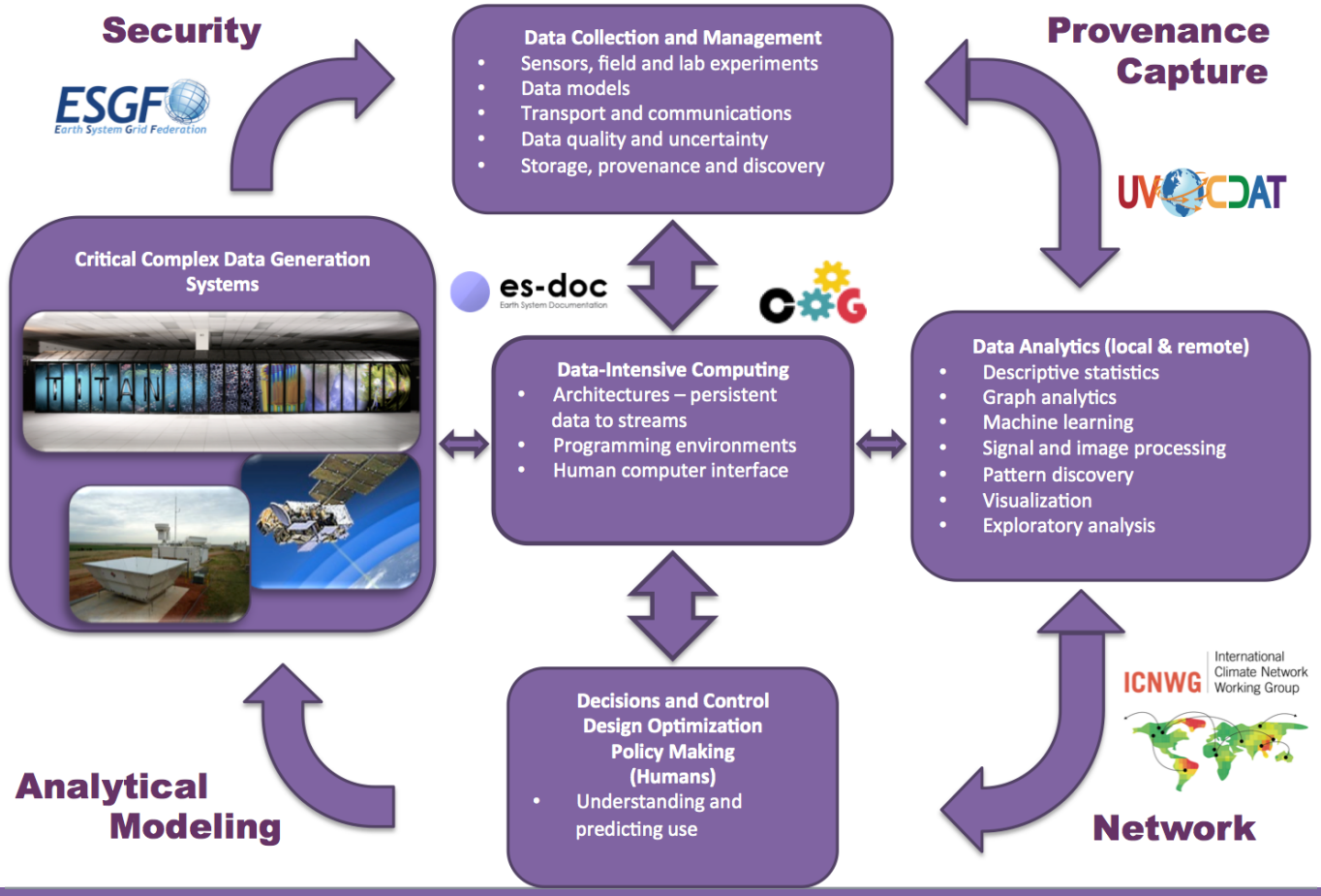
ESGF Governance Policy (http://esgf.llnl.gov/governance.html).

ESGF community activities: sub-tasks and task leaders

Sub-Task	Task Leads	Description
1. CoG User Interface Working Team	Cecelia DeLuca (NOAA) and Luca Cinquini (NOAA)	Improved ESGF search and data cart management and interface
2. Compute Working Team	Charles Doutriaux (DOE) and Daniel Duffy (NASA)	Developing the capability to enable data analytics within ESGF
3. Dashboard Working Team	Sandro Fiore (IS-ENES)	Statistics related to ESGF user metrics
4. Data Transfer Working Team	Lukasz Lacinski (DOE) and Rachana Ananthakrishnan	ESGF data transfer and enhancement of the web-based download
5. Documentation Working Team	Matthew Harris (DOE) and Sam Fries (DOE)	Document the use of the ESGF software stack
6. Identity Entitlement Access	Philp Kershaw (IS-ENES) and Rachana Ananthakrishnan (DOE)	ESGF X.509 certificate-based authentication and improved interface
7. Installation Working Team	Nicolas Carenton and Prashanth Dwarakanath (IS-ENES)	Installation of the components of the ESGF software stack
8. International Climate Network Working Group	Eli Dart (DOE/ESnet) and Mary Hester (DOE/ESnet)	Increase data transfer rates between the ESGF climate data centers
9. Metadata and Search Working Team	Luca Cinquini (NASA)	ESGF search engine based on Solr5; discoverable search metadata
10. Node Manager Working Team	Sasha Ames (DOE) and Prashanth Dwarakanath (IS-ENES)	Management of ESGF nodes and node communications
11. Provenance Capture Working Team	Bibi Raju (DOE)	ESGF provenance capture for reproducibility and repeatability
12. Publication Working Team	Sasha Ames (DOE) and Rachana Ananthakrishnan	Capability to publish data sets for CMIP and other projects to ESGF
13. Quality Control Working Team	Martina Stockhause (IS-ENES) and Katharina Berger (IS-ENES)	Integration of external information into the ESGF portal
14. Replication Working Team	Stephan Kindermann (IS-ENES) and Tobias Weigel (IS-ENES)	Replication tool for moving data from one ESGF center to another
15. Software Security Working Team	Prashanth Dwarakanath (IS-ENES) and Laura Carriere (NASA)	Security scans to identify vulnerabilities in the ESFF software
16. Tracking / Feedback Notification Working Team	Sasha Ames (DOE)	User and node notification of changed data in the ESGF ecosystem
17. User Support Working Team	Torsten Rathmann (IS-ENES) and Matthew Harris (DOE)	User frequently asked questions regarding ESGF and housed data
18. Versioning Working Team	Stephan Kindermann (IS-ENES) and Tobias Weigel (IS-ENES)	Versioning history of the ESGF published data sets

Further elaborations of the sub-tasks are described in the 2015 and 2016 ESGF progress reports and the ESGF Implementation Plan, which can be found online at http://esgf.llnl.gov/.

Data integration systems for climate science: data ecosystem and data flow



"Critical Complex Data Generation Systems" are housed and securely managed at many worldwide sites with the ESGF software stack. Data ecosystem, where provenance capture is pervasive throughout. Local and remote computation for more data-intensive and compute-intensive user requests. The network must be able to move petabytes of data between data centers. Finally, analytical modeling assists users in making smart choices in utilizing community resources for moving and computing large-scale data.

ESGF readiness for CMIP6

ESGF is systematically and rigorously developing, testing, evaluating, and documenting its subcomponents and associated tasks via teleconferences, face-to-face workshops and conferences, written reports, and journal publications[2].

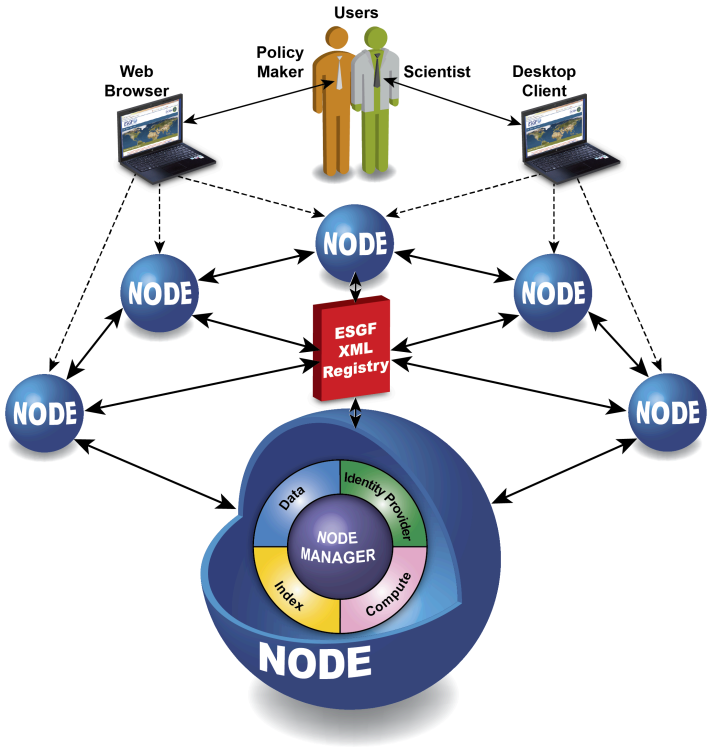
- ESGF Tier 1 and Tier 2 Node Requirements (under development);
- Data Storage and Replication Plan (under development);
- User Training Plan (under development); and
- ESGF CMIP6 Readiness Document (under development).

Living Document	Web Link <URL>	Description
Strategic Roadmap	http://esgf.llnl.gov/media/pdf/2015-ESGF-Strategic-Plan.pdf	This "Strategic Roadmap" describes the ESGF mission and an international integration strategy for data, database and computational architecture, and stable infrastructure highlighted by the ESGF Executive Committee. These highlights are key developments needed over the next five to seven years in response to large-scale national and international climate community projects that depend on ESGF for success.
Implementation Plan	http://esgf.llnl.gov/media/pdf/ESGF-Implementation-Plan-V1.0.pdf	The "Implementation Plan" describes how the ESGF data management system will be deployed, installed, and transitioned into an operational system. It contains an overview of the system, a brief description of the major tasks involved in the implementation, the overall resources needed to support the implementation effort (such as hardware, network, software, facilities, materials, and personnel), and any site-specific implementation requirements.
Software Security Plan	http://esgf.llnl.gov/media/pdf/ESGF-Software-Security-Plan-V1.0.pdf	The primary purpose of the "Software Security Plan" is to have a systematic approach to releasing a secure ESGF software stack (both major and minor ESGF releases) to the community. This is being done within the context of the ESGF Software Development Life Cycle (SDLC). This plan's emphasis is on the "release" phase of a typical SDLC and its prerequisites, which also depends upon development and maintenance (design and build) aspects of the SDLC.
Policies and Guidelines	http://esgf.llnl.gov/media/pdf/ESGF-Policies-and-Guidelines-V1.0.pdf	The ESGF is composed of groups and institutions that have elected to work together to operate a global infrastructure in support of climate science research. Although anyone is welcome to download, install, and run a copy of the ESGF software stack as a stand-alone node, joining the global federation requires understanding and abiding by the established ESGF policies and guidelines. This policies and guidelines are in place to provide the best possible experience to the community, ensure security and stability, and facilitate the job of the staff administering the ESGF nodes.
ESGF Root Certificate Authorities (CA) Policy & Certificate Practices Statement	http://docs.google.com/documents/d/16dxkvZy4J83j1nVL8vc_AwqGUSL52m-n8ulXU9eKhpQ/edit?usp=sharing	This document describes the set of rules and procedures established by the ESGF CA Policy Management Authority for the operation of the ESGF Root CA PKI services. ESGF PKI services. The Certificate Policy (CP) describes the requirements for operation of the PKI and for granting PKI credentials as well as lifetime management of those credentials. The Certificate Practices statement (CPS) describes the actual steps that ESGF takes to implement the CP. These two statements taken together are designed so that a Relying Party can look at them and obtain a understanding of the trustworthiness of credentials issued by the ESGF Root CA.

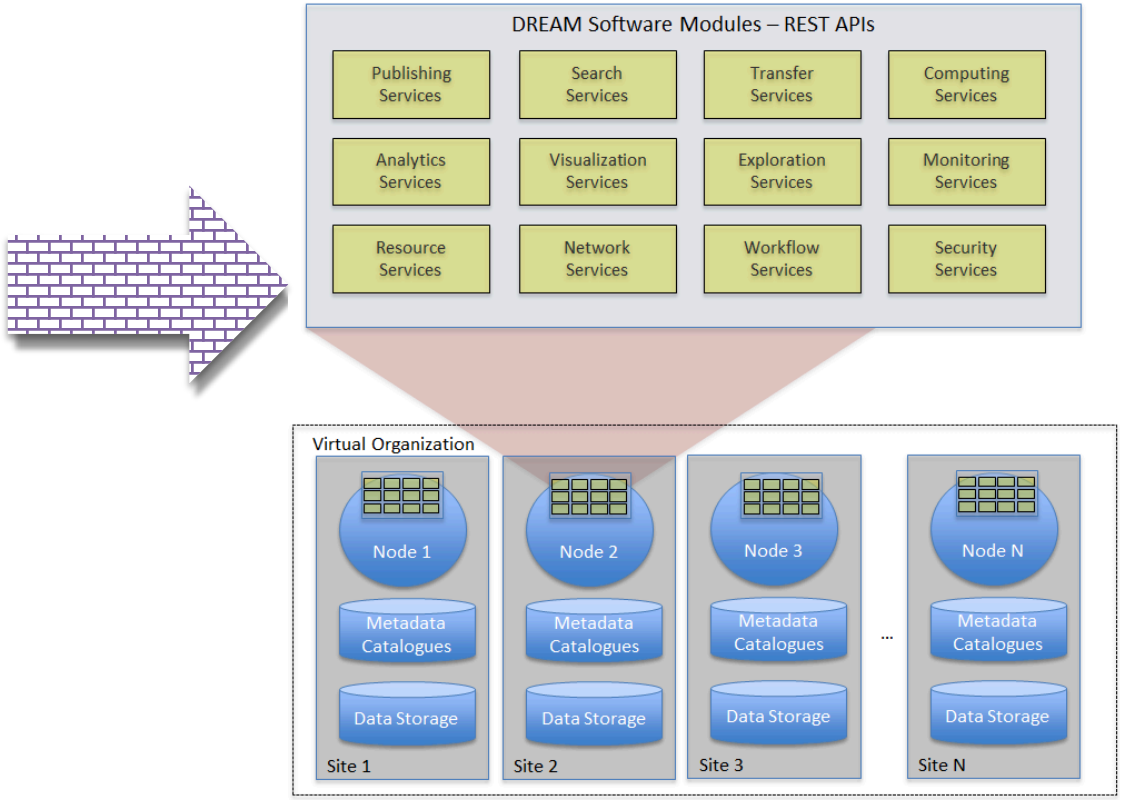
2. Dean N. Williams, V. Balaji, Luca Cinquini, Sébastien Denvil, Daniel Duffy, Ben Evans, Robert Ferraro, Rose Hansen, Michael Lautenschlager, and Claire Trenham, "A Global Repository for Planet-Sized Experiments and Observations", Bulletin of the American Meteorological Society, June 2016, doi: http://dx.doi.org/10.1175/BAMS-D-15-00132.1.

Next generation ESGF architecture by DREAM: modularization, composition, and RESTful APIs

ESGF Peer-to-Peer Architecture

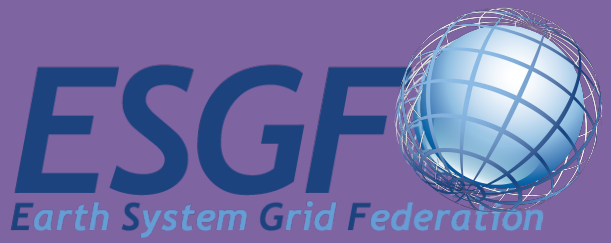


DREAM Template Architecture



High-level template architecture highlighting its modular structure (lower box) and its distributed topology (upper box). This architecture conforms to established standards across the federation and allows for a virtual computational environment.

- esgf.llnl.gov; ESGF public website
- esgf.llnl.gov/reports.html; ESGF reports
- github.com/esgf; software repository website
- icnwg.llnl.gov; international network website
- www.earthsystemcog.org/projects/cog/tutorials_web; CoG tutorial



Possible ESGF Tier 1 and Tier 2 node site requirements for CMIP6

- **Tier 1 sites**
  - Tier 1 sites are expected to run the full suite of ESGF services for data and user management, which can be used to support their own activities and those of Tier 2 sites
  - Have an uptime (>98%)
  - Have 10 petabytes of spinning disk storage space
  - Have at least a 10 gigabits per second connection to their wide-area network provider with plans to upgrade beyond 10Gbps by 2017
  - Run a 10 gigabits per second perfSONAR host (preferably on a physical server),
  - Deploy at least four 10 gigabits per second Data Transfer Nodes (DTNs) in a "Science DMZ" environment with plans to run production ESGF data services on the DTNs by 2017
  - Deploy sufficient high-performance storage to allow the DTNs to effectively serve CMIP6 data at high performance levels
  - Publish data using GridFTP and Globus URLs in addition to Wget URLs,
  - Configure the DTNs to use Globus as well as GridFTP and Wget, and
  - Use Synda for data replication between Tier 1 sites.
  - Core monitoring services
    - Unnoticed downtime by system administrator should not happen
    - Certificates end of validity monitoring
  - Quality of Service (TBD)

- **Tier2 sites**
  - Centers that typically have fewer physical or staff resources available for ESGF interactions but that still need to distribute a certain (possibly significant) amount of data to the scientific community.
  - Tier 2 sites are encouraged to leverage some of the services supported by Tier 1 sites, such as a Metadata Index and Identity Provider, and focus instead on supporting local services for data download and possibly analysis.

The full ESGF Tier 1 and Tier 2 Node requirements are under development. Decisions cannot be made until after the estimated size of the CMIP6 archive is determined.